06-07 00

A

KAPLAN & GILMAN, L.L.P. Counselors at Law

900 Route 9 North
Woodbridge, NJ 07095
Telephone (732) 634-7634
Facsimile (732) 634-6887
www.kaplangilman.com

Patents
Trademarks
Copyrights
Licensing
Trade Secrets

Date: June 6, 2000

Re: Inventor(s): Chris A. Hamilton

Title: IMPROVED CONTROL OF VIDEOCONFERENCING USING ACTIVITY DETECTION

Atty. Docket No.: 024/1

Box NEW APP. FEE
Assistant Commissioner for Patents
Washington, D.C. 20231

Dear Sir:

Submitted herewith is the above-identified continuation patent application. Please abandon the parent case (Serial No. 09/098,911) only after a filing date is granted to this case. Also enclosed are:

1) A self-addressed, stamped return postcard;

2) A Petition for Three-Month Extension of Time;

3) 3 sheets of informal drawings of Figs. 1-3;

4) A check in the amount of $690.00 made payable to the "Commissioner of Patents and Trademarks"; and

5) A Preliminary Amendment.

Respectfully submitted,

KAPLAN & GILMAN, L.L.P.

Jeffrey I. Kaplan
Reg. No. 34,356

JIK/pa
Enclosures

CERTIFICATE OF MAILING

Express Mail mailing label number: EL559055229US
Date of Deposit: June 6, 2000

I hereby certify that this paper or fee is being deposited with the United States Postal Service *Express Mail Post Office to Addressee* service under 37 C.F.R §1.10 on the date indicated above and is addressed to Box NEW APP. FEE, Assistant Commissioner for Patents, Washington, D.C. 20231

Paula M. Halsey
(Typed or printed name of person mailing paper or fee)

Paula M. Halsey
(Signature of person mailing paper or fee)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	**Chris A. Hamilton**
Title of Invention	:	**IMPROVED CONTROL OF VIDEOCONFERENCING USING ACTIVITY DETECTION**

Assistant Commissioner for Patents
Washington, DC 20231

PRELIMINARY AMENDMENT

SIR:

This application is a continuation of U.S. Patent Application Serial No. 09/098,911. The six-month date for response to the outstanding Office Action expires today, and a Petition for a Three-Month Extension of Time is enclosed herewith.

All of the claims have been amended. The newly amended claims call for the speaker to be identified by the use of a combination of image processing and voice activity detection. This technique is fully described at page 5 of the present application.

Applicant is in the process of translating the Japanese prior art relied upon in the parent case. The abstract portion, which is in English, does not appear to disclose this technique and in fact, specifically indicates that the presence or absence of speech is identified "according to the voice level of the conference participant." Such a technique could fail based upon loud ambient noise, but applicant's invention of utilizing the additional step of image processing with speech activity detection prevents such a failure.

Applicant submits that it would not in any way be obvious to combine the teachings of the Ogata reference with Zhou, U.S. Patent No. 5,512,939. Specifically, Zhou dynamically allocates image encoding bandwidth to different portions of the signal, increasing the resolution of the lips

when the person is speaking. However, Zhou in no way deals with videoconferencing or with distinguishing among various conference participants with different lip movements. Instead, Zhou simply enhances portions of the video screen with the moving lips. It is submitted that there is no suggestion in either of the references to change the display and designate the speaker by combining both an audio and video signal to drive a video generator to generate a border around the speaker's image. This combination is neither suggested nor disclosed in either of the references at issue, and it is submitted that applicant's claims are thus patentable.

The Examiner is authorized to deduct any additional fees believed due from Deposit Account No. 11-0223.

Respectfully submitted,

KAPLAN & GILMAN, L.L.P.
900 Route 9 North
Woodbridge, NJ 07095
Telephone (732) 634-7634

DATED: June 6, 2000

Jeffrey J. Kaplan, Esq.
Reg. No. 34,356

IMPROVED CONTROL OF VIDEO CONFERENCING USING ACTIVITY DETECTION

RELATED APPLICATION

This is a continuation of application Serial No. 09/098,911.

TECHNICAL FIELD

This invention relates to video conferencing, and more specifically, to an improved technique of allowing various members of a video conference to identify which subset of a plurality of conference members are speaking at any time. In particular embodiments, items such as voice activity detection and image recognition software are used to automatically determine which of the conference members are speaking.

BACKGROUND OF INVENTION

Video conferencing is a technique utilized in order to provide both video and audio information from one or more users to a plurality of other users. Typically, a conference bridge is utilized to connect several participants of the video conference, and the signal received at the conference bridge from each conferee is broadcast to the other conference members. As a conferee uses the conference station, he/she views separate images from each of the other conference stations. Figure 2 shows an example of a conference station as viewed by a conferee participating in a conference with four other conferees. As seen in Figure 2, the video information from each of the four other conferees is displayed on a conference station video monitor, usually a personal computer. In this example, conferee 2 is missing, since it is the conference station of conferee 2 being viewed. Of course, a conferee may choose to see his own image on the screen.

Recently, much of the available conferencing technology is becoming focused on digital

techniques. More specifically, with the availability of Internet access becoming less expensive and more widespread, it has become possible to implement the video conferences over the Internet or other similar data networks. Implementation of such conferences in the digital domain provides improved clarity, availability of compression techniques, etc. Additionally, with the price of personal computers getting lower and the speed of such computers increasing, it is possible to very inexpensively implement functions such as speech recognition, image processing, etc. Little advantage has been taken of the additional capabilities available in PC-based conference stations, and more particularly, of the ability of such conference stations to provide advanced signal processing functions.

There has been little research to date focused upon taking advantage of the additional capabilities of implementing video conferencing in the digital domain. Specifically, effective techniques which may reduce the confusion as to which participants in a video conference are speaking are not found in the prior art. In addition, the prior art does not utilize the combination of video and audio information for the purpose of voice activity detection.

SUMMARY OF THE INVENTION

The above and other problems of the prior art are overcome in accordance with the present invention which relates to an improved video conferencing system which provides for a technique of informing video conference members which subset of conference members are speaking at any given time. Technologies utilized include voice activity detection, speaker identification, and image recognition, or other such items.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a conceptual diagram of a video conferencing arrangement;

Figure 2 depicts an exemplary video screen showing four conferees; and

Figure 3 depicts a slightly more detailed diagram of a conference bridge for use with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Fig. 1 shows a typical conference bridge arrangement for implementing a multi-party video conference. The arrangement shown in Fig. 1 includes a conference bridge 108 and a plurality of conference stations 102-107. The conference bridge 108 is implemented from well-known off the shelf components. The conference bridge 108 receives video signals from conference stations 102-107, and transmits to each conference station a signal indicative of the video and audio from all of the other conferencing stations.

In accordance with one embodiment of the present invention, a video conference speaker identification subsystem is utilized at each conference station 102-107 in order to allow conferees to more easily determine which of the conference members may be speaking at any time. Specifically, if the speaker identification subsystem located at, for example, conference station 102, detects or is informed that the person at conference station 103 is talking, then conference station 102 may act automatically in such a way as to inform the user at station 102 of this fact. Typically in multi-party conferences, a user will be presented at his conference station with the images of each of the other conference members as well as an monophonic mix of the audio source from each of the

other conference stations. If many images are present on the conference station screen, then it may not be apparent who is speaking without a careful visual search of the screen images. In the present example, the conference member at station 102 would be presented with textual or graphic or other information informing him that the conferee at station 103 was speaking.

In one exemplary implementation, a voice activity detector is utilized in order to determine which of the conferees may be speaking at any time. Voice activity detectors are well known in the field of telecommunications and in the present invention could be implemented at the conference station or at the conference bridge server. In either case it would then be possible for the conference system to be able to differentiate those conferees who are speaking from those who are not. This differentiation can be useful. For example, the screen images of speaking conferees could be altered. Thus, for example, a border could be drawn around the image of any party speaking indicating to the other conferees that this image is the source of speech. Referring to Figure 2, if conferee 201 begins speaking such that his voice is significantly louder than the other conferees, a bright border would appear around the image of conferee 201.

In an additional implementation, an improved voice activity detector (VAD) is utilized in order to determine which of the conferees may be speaking at any time. This improved VAD makes use of the audio signal as well as the video signal transmitted by a conference station. In particular, a traditional VAD is combined with image analysis and recognition software to improve the accuracy of the VAD. Image analysis and recognition techniques are well known in the field of image processing and may be employed here to analyze the image of a conference member to: (1) recognize the lips of within the image of the conferee and (2) to determine if the lips of the conferee are moving in a way that is reasonably consistent with the audio signal transmitted by the conference

station. Thus, voice activity is detected when both audio and video components of the outgoing conference signal are consistent with human speech. Knowledge of such activity can be useful not only in allowing others within the conference to know which members are speaking, but also to save network bandwidth, etc.

Figure 3 shows a slightly more detailed embodiment of the present invention comprising a plurality of receiving modules 301-303 and transmission modules 304-306. The exemplary simplistic arrangement of Figure 3 is intended to conference three video conference stations together, with each transmission module 304-306 conveying to a conference station the two other conference station signals. Control lines 307-309 serve to activate and deactivate the functions previously discussed. For example, if it is determined that the received video stream from the conference station 301 is to be surrounded with a particular border, control line 307 instructs bridging hardware 310 appropriately. The bridging hardware 310 will then insert the border prior to placing the combined image for transmission on the appropriate two transmission modules 304-306.

The above describes the preferred embodiments of the invention. Various modifications and additions will be apparent to those of skill in the art.

WHAT IS CLAIMED:

1. An improved video conferencing system comprising:

a switch for interconnecting a plurality of video conference stations;

a processor for visually altering an image of at least one of a plurality of remotely located conferees that is speaking at a particular time, said processor including an algorithm for determining which speakers are speaking by comparing visual lip movement with an audio signal.

2. Apparatus of claim 1 wherein said processor is responsive to a voice activity detector located at remote conference stations.

3. Apparatus of claim 2 wherein said processor comprises software for highlighting a border around an image of a remote speaker determined to be speaking.

4. A video conference station comprising:

a transmitter to transmit a combined audio video signal to a video conference bridge; and

a processor within said video conference station for performing voice activity detection and image analysis to determine when a conferee located at said video conference station is speaking.

ABSTRACT

A video conferencing system employs a plurality of video conferencing stations, each of which includes a voice activity detector. When the images of remote conferees are displayed on a video conference station, the voice activity detection is utilized in order to designate which remote speaker is presently speaking.

FIG. 1

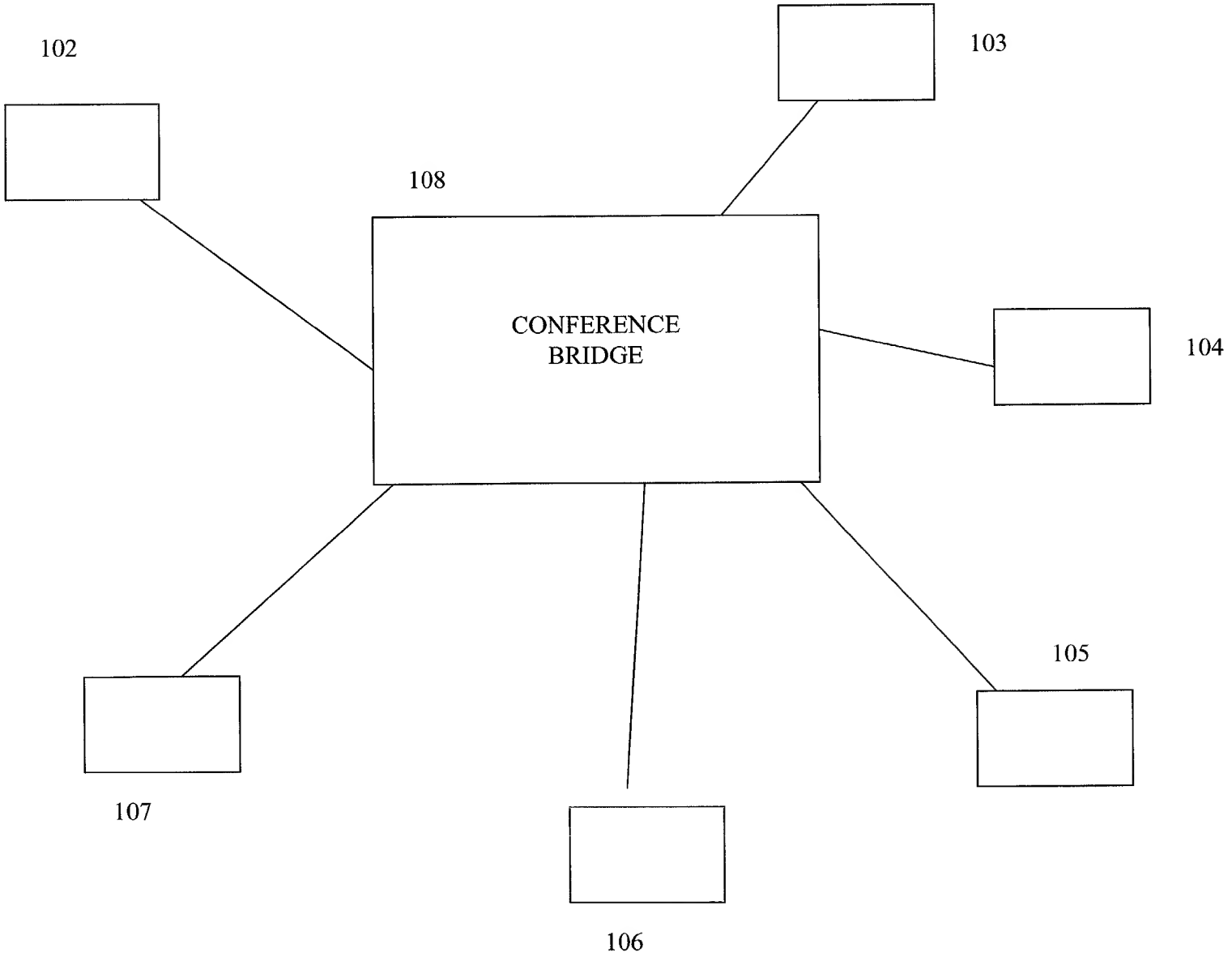


FIG. 2

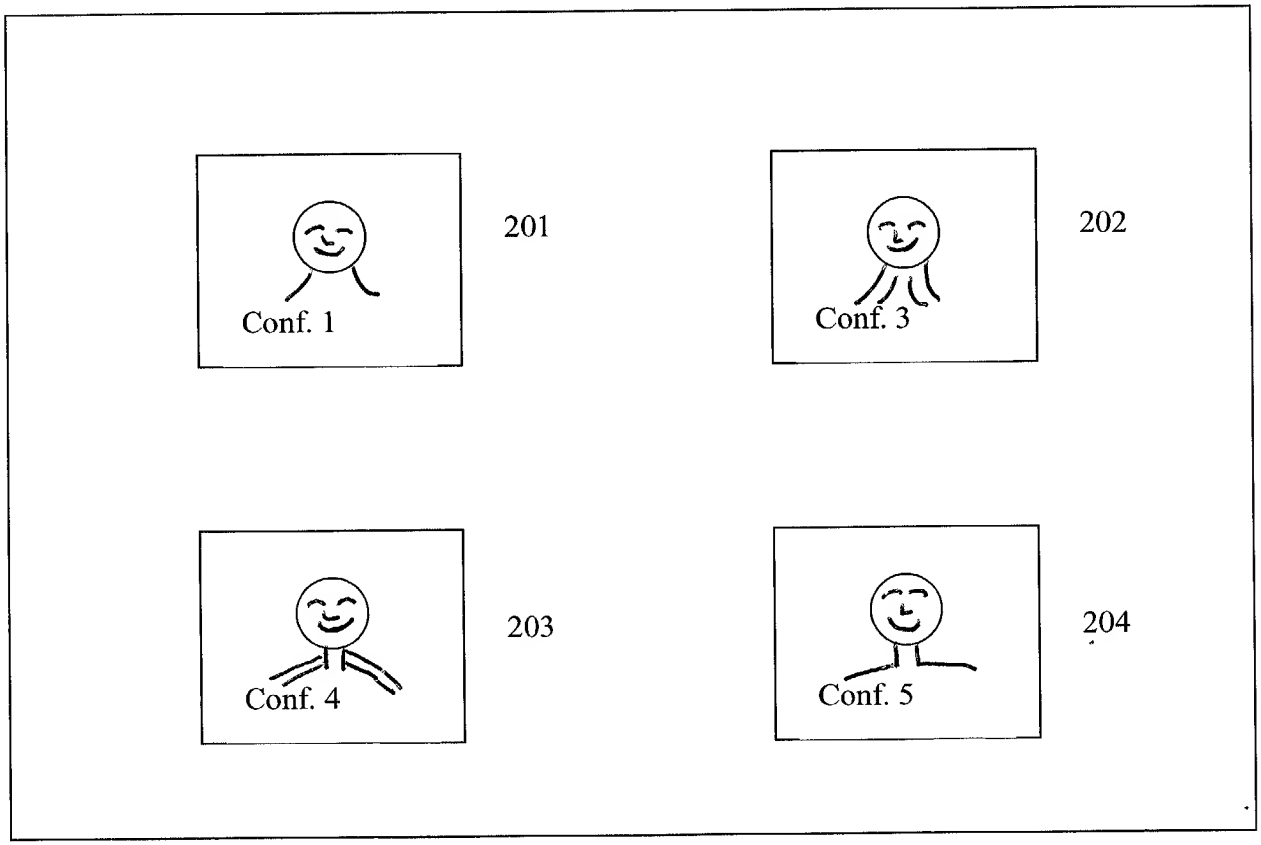
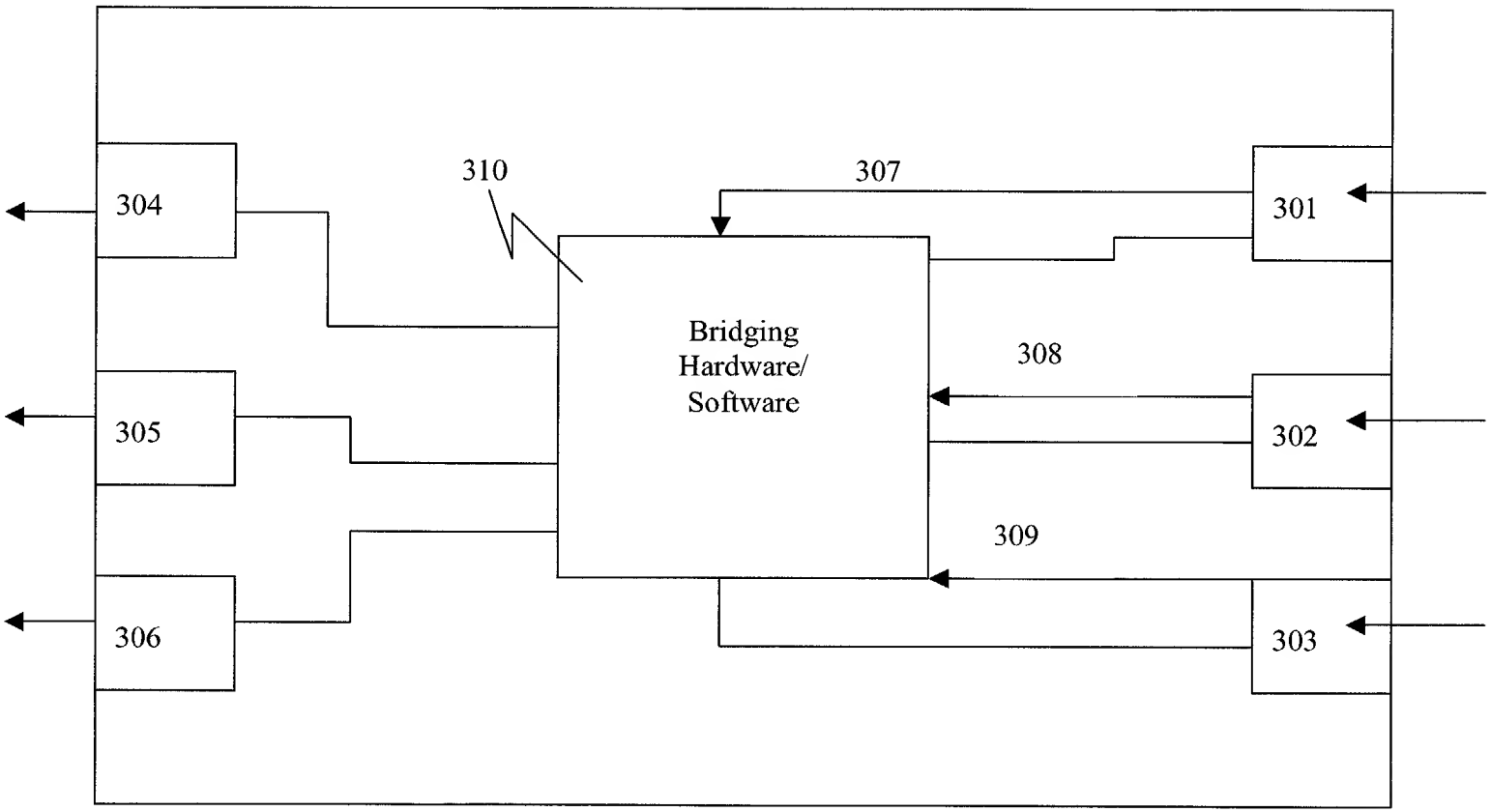


FIG. 3


304
305
306
310
307
Bridging
Hardware/
Software
308
309
301
302
303